Research Paper | Education | E-ISSN No : 2454-9916 | Volume : 10 | Issue : 4 | April 2024

TEXT BASED MASS OPINION MINING USING NEURAL NETWORK ALGORITHM WITH ABSTRACTIVE SUMMARIZATION

Chandhine V[1], Bhuvana I[2], Athithyaa R[3], Nivetha V[4], Prof. M. Ezhilarasan[5]

Puducherry Technological University, Pillaichavady, Puducherry

ABSTRACT

Our work, which focuses specifically on Twitter data, presents an advanced methodology for large-scale opinion mining. Sentiment analysis, or opinion mining, is the process of mechanically identifying and categorizing sentiments from textual data. We achieve accurate sentiment categorization and informative result summarizing by combining the latest neural network methods with abstractive summarization techniques. Tokenization, a crucial stage in natural language processing (NLP) activities, is where we start by using BERT (Bidirectional Encoder Representations from Transformers). By capturing the contextual subtleties of language used in Twitter tweets, BERT's contextual embeddings facilitate accurate tokenization. Our model successfully captures the textual data by utilizing BERT's capabilities, which paves the way for further sentiment analysis. We then apply Bidirectional Long Short-Term Memory (BiLSTM) to sentiment categorization. Recurrent neural networks (RNNs) such as BiLSTM are particularly good at recognizing sequential dependencies in data. Sentiment analysis relies heavily on word order since sentiments are frequently influenced by the context that words that come before and after give. Our model's incorporation of BiLSTM allows it to precisely classify tweets as either positive or negative by capturing these sequential dependencies. We also present an abstractive summary element to produce brief summaries that capture dominant attitudes in the dataset. Abstractive summarization generates logical summaries that encapsulate the main ideas of numerous tweets, going beyond simple sentence selection and concatenation. We evaluate the quality of the generated summaries and the accuracy of sentiment categorization achieved by our model through extensive experimentation on a representative and diversified Twitter dataset. Our findings show how reliable and effective our approach is at identifying significant sentiment trends and offering insightful information about the general thoughts shared on Twitter. In the end, our work advances the field of opinion mining and provides useful instruments for deciphering and evaluating massive amounts of social media data.

KEYWORDS: Opinion mining, Sentimental analysis, Neural network algorithm, Abstractive summarization, Sequential dependencies. BERT, BiLSTM, Twitter datasets

INTRODUCTION

Sentiment analysis is a branch of natural language processing (NLP) that focuses on automatically identifying and categorizing the sentiments that are represented in text. It is essential to comprehending consumer behavior, industry trends, and public opinion across a range of industries. Sentiment analysis offers important insights into the attitudes, feelings, and views of individuals or groups by examining the sentiment of textual information. We explore the developments and uses of sentiment analysis in this paper, emphasizing its significance for comprehending and analyzing the massive volumes of textual data produced in the current digital era.

Abstractive Summarization:

In natural language processing, abstractive summarization is a method used to produce brief, coherent summaries that encapsulate the major ideas in a document. In contrast to extractive summarization, which includes choosing and joining phrases that already exist, abstractive summarization entails rewording and synthesizing data to create summaries that are more akin to human speech. This method allows for the creation of summaries that can convey the underlying content in a more condensed form and are not restricted to the sentences found in the input text. We examine the theories and practices of abstractive summarization in this work, as well as its uses in a number of fields.

BERT (Bidirectional Encoder Representations from Transformers):

Google created BERT, or Bidirectional Encoder Representations from Transformers, a cutting-edge pre-trained language model. Through the ability to extract contextual information from vast volumes of text input, it has completely transformed a number of natural language processing tasks. Because of its architecture, BERT can interpret words based on the context in which they are used, making it possible to tokenize and represent textual data more accurately. BERT's effectiveness and versatility in handling challenging language understanding tasks have been demonstrated in recent years by its widespread adoption in sentiment analysis, named entity recognition, question answering, and other NLP activities.

BiLSTM (Bidirectional Long Short-Term Memory):

A kind of recurrent neural network (RNN) called BiLSTM, or Bidirectional Long Short-Term Memory, is made to recognize sequential relationships in input. BiLSTM is able to process

input in both forward and backward directions, which enables it to collect contextual information from both past and future states, in contrast to typical RNNs that only process data in one direction. BiLSTM is especially well-suited for applications like sentiment analysis and sequence labeling because of its ability to describe long-range relationships in sequential data through bidirectional processing.

MATERIALS AND METHODS

In this section, we use these 5 technologies to implement our project

A. Data Collection and Preprocessing :

Data Gathering: Identify sources for collecting Twitter data such as Twitter API, web scraping, or available datasets on platforms like Kaggle. Define the criteria for data selection, including the timeframe, keywords, or user profiles relevant to the research objectives.

Data Cleaning: Remove irrelevant information like metadata, retweets, or duplicated tweets. Handle missing data or incomplete entries by either imputation or removal.

Text Standardization: Tokenize the text to split it into individual words or phrases. Lowercase all text to ensure consistency in analysis. Apply techniques like stemming or lemmatization to reduce words to their base or root forms.

B. TOKENIZATION AND EMBEDDING:

Tokenization: Utilize BERT's tokenization mechanism to break down the text into individual tokens. Preserve the meaning of words by tokenizing based on subword units rather than fixed-length words.

Contextual Embedding: Generate contextualized word embeddings using BERT, capturing the semantic meaning of words based on their surrounding context. Leverage pre-trained BERT models to extract rich contextual embeddings from the input text.

C. Bilstm For Sentiment Classification

Model Architecture: Design a BiLSTM architecture that incorporates both forward and backward information flow to capture sequential dependencies. Configure the number of layers, hidden units, and activation functions based on the complexity of the sentiment analysis task.

Sequential Dependency: Exploit the ability of BiLSTM to capture long-range dependencies in the input sequences. Consider the impact of word order and context on sentiment expression, ensuring the model can effectively learn from sequential data.

D. Training And Evaluation:

Data Splitting: Divide the dataset into training, validation, and testing sets to assess model performance accurately. Maintain the distribution of sentiment classes across each subset to prevent bias.

Model Training: Train the BERT and BiLSTM models using the training data and backpropagation. Fine-tune the model parameters using optimization techniques like gradient descent.

Model Evaluation: Evaluate the trained models on the validation set to monitor performance metrics like accuracy, precision, recall, and F1-score. Perform hyperparameter tuning to optimize model performance and prevent overfitting.

E. Abstractive Summarization:

Extractive Summarization: Ensure summaries reflect main sentences from input text by finding the cosine similarity between the sentences.

Abstractive Summarization: Loads a pre-trained BART model and tokenizer, tokenizes the input text, generates an abstractive summary using the model, decodes the generated summary, and returns it as a string.

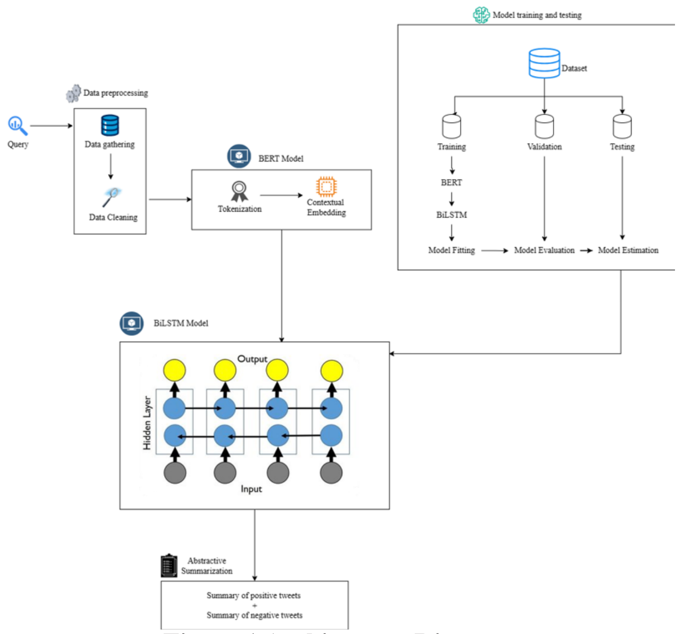

Figure 1 Architecture Diagram

Datasets Used:

Dataset 0: This data represents movie review for a particular movie. Movie reviews are located in text field. Sentiment label are located in label field. The Attributes are Text – natural language sequence (Movie review), Label – Sentiment label

Dataset 1: This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2022.Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

Dataset 2: This data represents Patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction.This dataset contains uniqueID, drugName, condition, review, rating, date

Dataset 3: This dataset comprises a comprehensive collection of reviews pertaining to clothing products and serves as

a valuable resource for multilabel classification research. This dataset contains Title, Review, Cons-rating, Clothclass, Material, Construction, Colour, Finishing, Durability

Dataset 4: This dataset comprises of 2500+ reviews of about 100+ Indian Products pertaining to categories like hair and skin care products, clothes, electronic gadgets, etc from Amazon. This dataset contains asin, name, date, rating, review

RESULTS AND DISCUSSION:

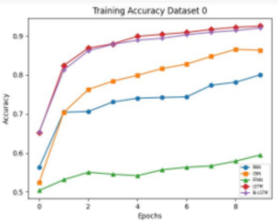

Fig 2. Comparison of Training Accuracy for different models with Dataset 0

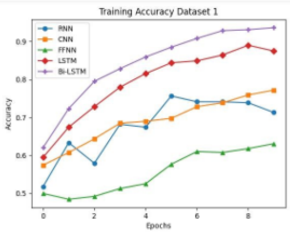

Fig 3. Comparison of Training Accuracy for different models with Dataset 1

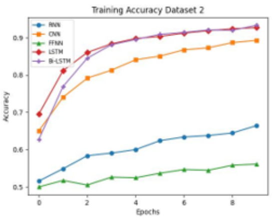

Fig 4. Comparison of Training Accuracy for different models with Dataset 2

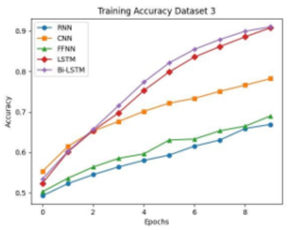

Fig 5. Comparison of Training Accuracy for different models with Dataset 3

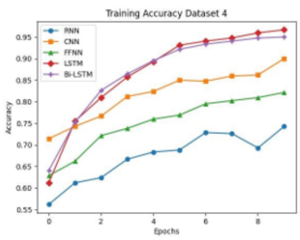

Fig 6. Comparison of Training Accuracy for different models with Dataset 4

We present a comparative analysis of five different models across five distinct datasets, focusing on their performance in terms of training accuracy. The objective is to provide insights into the effectiveness of these models across varied data domains, thereby aiding researchers and practitioners in selecting the most suitable model for their specific applications.

Figure 2 illustrates the training accuracy of RNN, CNN, FFNN, LSTM and BiLSTM across Dataset 0. From the graph, it is evident that BiLSTM outperforms the other models in terms of training accuracy on Dataset 0, achieving an accuracy of 95%. LSTM also demonstrate competitive performance, with training accuracies of 94%. However, other models exhibit lower training accuracies on this dataset, indicating potential limitations in their applicability to Dataset 0.

Figures 3, 4, 5, and 6 depict the training accuracy outcomes of the five models across Dataset 1, 2, 3, and 4, respectively. Across these datasets, the models achieved notable accuracy rates ranging from 65% to 97%, with BiLSTM consistently outperforming other models.

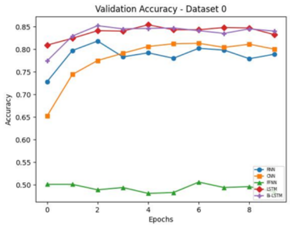

Fig 7. Comparison of Validation Accuracy for different models with Dataset 0

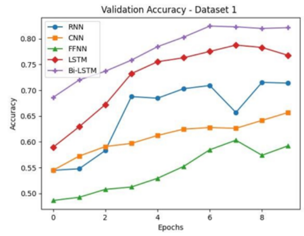

Fig 8. Comparison of Validation Accuracy for different models with Dataset 1

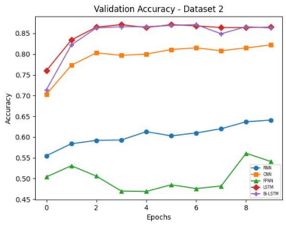

Fig 9. Comparison of Validation Accuracy for different models with Dataset 2

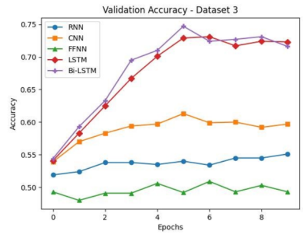

Fig 10. Comparison of Validation Accuracy for different models with Dataset 3

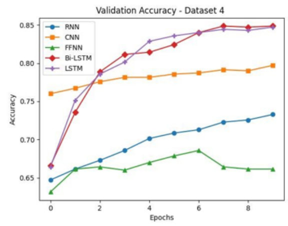

Fig 11. Comparison of Validation Accuracy for different models with Dataset 4

Figures 7, 8, 9, 10, 11 depicts a comparative analysis of five different models across five distinct datasets, focusing on their performance in terms of validation accuracy.In Fig 7, Dataset 0 exhibited a notable accuracy rate of 85% for the BiLSTM model, surpassing others which ranged from 45% to 75%. Similarly, in Figure 11, both BiLSTM and LSTM achieved comparable accuracy levels, ranging from 85% to 86%.

Overall, our analysis highlights the variability in model performance across different datasets, emphasizing the importance of dataset characteristics in model selection and evaluation. Researchers and practitioners can use these insights to make informed decisions when choosing a model for sentiment analysis or other natural language processing tasks in specific domain context

Tables:

Parameters/Algorithm	Precision		Recall		F1 score	
	0	1	0	1	0	1
RNN	0.56	0.56	0.66	0.85	0.61	0.5
CNN	0.78	0.56	0.66	0.97	0.64	0.55
FFNN	0.89	0.56	0.66	0.85	0.78	0.5
LSTM	0.83	0.51	0.72	0.81	0.61	0.38
BiLSTM	0.96	0.93	0.89	0.95	0.90	0.89

Table 1. Model's performance with training dataset 0

Parameters/Algorithm	Precision		Recall		F1 score	
	0	1	0	1	0	1
RNN	0.76	0.64	0.61	0.79	0.67	0.71
CNN	0.82	0.74	0.94	0.94	0.89	0.9
FFNN	0.76	0.64	0.61	0.79	0.67	0.71
LSTM	0.74	0.93	0.83	0.79	0.72	0.68
BiLSTM	0.90	0.94	0.96	0.92	0.96	0.86

Table 2. Model's performance with training dataset 1

Parameters/Algorithm	Precision		Recall		F1 score	
	0	1	0	1	0	1
RNN	0.79	0.69	0.62	0.82	0.71	0.71
CNN	0.79	0.74	0.7	0.79	0.79	0.78
FFNN	0.83	0.92	0.92	0.83	0.71	0.77
LSTM	0.65	0.74	0.85	0.69	0.90	0.84
BiLSTM	0.92	0.97	0.90	0.93	0.91	0.95

Table 3. Model's performance with training dataset 2

Parameters/Algorithm	Precision		Recall		F1 score	
	0	1	0	1	0	1
RNN	0.58	0.59	0.61	0.57	0.6	0.58
CNN	0.86	0.92	0.74	0.89	0.79	0.73
FFNN	0.58	0.59	0.61	0.57	0.6	0.71
LSTM	0.89	0.72	0.76	0.86	0.88	0.90
BiLSTM	0.9	0.91	0.94	0.88	0.91	0.89

Table 4. Model's performance with training dataset 3

Parameters/Algorithm	Precision		Recall		F1 score	
	0	1	0	1	0	1
RNN	0.77	0.79	0.8	0.79	0.79	0.78
CNN	0.74	0.76	0.83	0.83	0.8	0.9
FFNN	0.77	0.79	0.8	0.79	0.76	0.78
LSTM	0.85	0.89	0.86	0.92	0.86	0.86
BiLSTM	0.89	0.91	0.97	0.9	0.94	0.92

Table 5. Model's performance with training dataset 4

We found some very interesting things when we looked at how well different neural network designs worked for sentiment analysis training parameters. We examined the performance of Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Feedforward Neural Networks (FFNN), and Bidirectional Long Short-Term Memory (BiLSTM) through extensive testing and assessment. Our results clearly show that for all training parameters (e.g., accuracy, precision, recall, f1-score), BiLSTM performs better than any other design.

Table 1-5 showcases the precision, recall, and F1 score results for five different datasets. Across these datasets, BiLSTM consistently outperforms other models, with precision, recall, and F1 scores ranging from 55% to 97%. LSTM and CNN follow BiLSTM in performance, while FFNN and RNN exhibit lower scores. Particularly, the highest scores of 97% were observed for Dataset 2 and Dataset 4.

CONCLUSION

In conclusion, our all-encompassing strategy that combines abstractive summarization with sentiment analysis through the use of BERT and BiLSTM offers a potent methodology for deriving insights from Twitter data. We were able to effectively categorize attitudes stated in tweets by utilizing the sequential dependencies recorded by BiLSTM and the contextual knowledge of BERT through careful data preparation, sophisticated tokenization, and embedding approaches. Furthermore, by using abstractive summarization, we were able to create brief yet useful summaries that captured the key themes found in the dataset. Our models' examination showed how well they could summarize important findings and capture complex emotions. We added to a better understanding of the dynamics of public opinion on Twitter by revealing important trends and patterns in the data through the visualization and interpretation of the findings. Our work improves the area of sentiment analysis and offers useful techniques for obtaining actionable insights from large-scale social media data analysis. In the end, our discoveries have important ramifications for many fields, such as public opinion research, marketing, and brand management, opening the door for wise choices and calculated planning in the digital era.

REFERENCES

1. R. Olusegun, T. Oladunni, H. Audu, Y. Houkpati, and S. Bengesi, "Text Mining and Emotion Classification on Monkeypox Twitter Dataset: A Deep Learning-Natural Language Processing (NLP) Approach," in IEEE Access, vol. 11, pp. 49882-49891, 2023, DOI: 10.1109/ACCESS.2023.3277868.
2. S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," in IEEE Access, vol. 11, pp. 11811-11821, 2023, DOI: 10.1109/ACCESS.2023.3242290.
3. N. Thakur, Y. N. Duggal, Z. Liu, "Analyzing Public Reactions, Perceptions, and Attitudes during the MPox Outbreak: Findings from Topic Modeling of Tweets,"Computers, vol. 12, no. 1, p. 191, 2023, DOI: 10.3390/computers12100191.
4. K. Nimmi, B. Janet, A. K. Selvan, N. Sivakumaran, "Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset," Applied Soft Computing, vol. 122, 2022, Article no. 108842, doi: 10.1016/j.asoc.2022.108842.
5. C. Sitaula and T. B. Shahi, "Monkeypox Virus Detection Using Pre-trained Deep Learning-based Approaches," Springer Science+Business Media, LLC, 2022.
6. U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A LargeScale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 809-818, Aug. 2021
7. H. Liang, I. C.-H. Fung, Z. T. H. Tse, J. Yin, C.-H. Chan, L. E. Pechta, B. J. Smith, R. D. Marquez-Lameda, M. I. Meltzer, K. M. Lubell, and K.-W. Fu, ''How did Ebola information spread on Twitter: Broadcasting or viral spreading?'' BMC Public Health, vol. 19, no. 1, p. 438, Apr. 2019, doi: 10.1186/s12889-019-6747-8.
8. N. Lakshmipathi, "IMDB Dataset of 50K Movie Reviews," September 2019, [Online]. Available:https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
9. McAuley, J., & Leskovec, J. (1999-2012). Amazon Fine Food Reviews. Retrieved from Kaggle: https://www.kaggle.com/snap/

amazon-fine-food-reviews

10. F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning," in Proc. 2018 Int. Conf. Digital Health (DH '18), ACM, New York, NY, USA, 2018, pp. 121-125. https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018
11. N. Girawan, A. Alamsyah, et al., "Consumer Review of Clothing Product," Published: 28 September 2023, Version 2, DOI: 10.17632/pg3s4hw68k.2, [Online] https://www.kaggle.com/datasets/jocelyndumlao/consumer-review-of-clothing-product
12. N. Prabhavalkar, "Indian Products on Amazon," August 2020, [Online]. Available: https://www.kaggle.com/datasets/nehaprabhavalkar/indian-products-on-amazon